Chapter 16
Identifying the Reaction Mechanisms of Inteins with QM/MM Multiscale Methods

Philip T. Shemella and Saroj K. Nayak

Abstract With a series of quantum mechanical calculations ranging from gas phase, to an implicit solvent scheme, to combined quantum/classical simulations, we have provided insight into some of the key steps of intein reactions. These studies may be exploited for many applications involving inteins including molecular
AQ1 switches and sensors as well as controlled drug delivery.

1 Introduction

1.1 Computational Background

Nestled between experiment and pure theory, computational chemistry has become an integral tool for researchers working in physics, chemistry, and biology, as well as nanotechnology and biotechnology. Computer simulations allow the researcher to access states both visible and invisible to experiment, and make predictions based on this knowledge. A chemical reaction may be quantified by the amount of reactants, the amount of products, and the time elapsed. To explain a mechanism and molecular structure and energies on the atomic level, computational methods are important.

The field of computational chemistry spans many length and time scales. To simulate protein folding, which requires an extremely long simulation trajectory, amino acids may be "coarse-grained," where the atomic description of each side chain is aggregated into a composite value. To achieve long trajectories this approximation as well as others are essential. However, to calculate the pK_a of a side chain or the chemical shifts *via* nuclear magnetic resonance (NMR), not only will an atomic level description be necessary, but also a method that can calculate observable properties from first principles is often required.

AQ2 P.T. Shemella and S.K. Nayak
Rensselaer Polytechnic Institute, 110 8th St., Troy, NY 12180, USA
e-mail: nayaks@rpi.edu

S. De et al. (eds.), *Computational Modeling in Biomechanics*,
DOI 10.1007/978-90-481-3575-2_16, © Springer Science+Business Media B.V. 2010

The energies associated with bond breakage and formation are an essential property for an enzymatic processes. For example, a change in energy barrier of ~1.4 kcal/mol at room temperature corresponds to an order of magnitude change in the reaction rate. States observed at equilibrium may be predicted based on relative energies between structures. To computationally access the energy of the system, and to do so not only for equilibrium structures but also for transition states, first principles electronic structure calculations are required. Using an all-electron method, the electron orbitals are considered variable and flexible, and they depend on neighboring atoms and environment. This is important because the chemistry at transition states may vary greatly from equilibrium structures: instead of four bonds, carbon atoms may have three or five bonds during a chemical reaction. Transition states are where quantum mechanical principles dominate. By solving the Schrödinger equation for all electrons, and relaxing their orbital positions and therefore allowing the electron density to vary, an accurate description of the system can be obtained that is useful for understanding fundamental chemistry both near and far from equilibrium.

1.2 Inteins Background

Protein splicing involves the autocatalytic release of a peptide segment, termed an intein, with the joining of two flanking protein sequences (exteins) [1, 2]. Inteins are autocatalytic proteins that exist in all three domains of life. Experiments have identified key reaction steps in protein splicing whereas sequence comparisons have revealed the conserved amino acids required for this reaction. Figure 1 shows a schematic for conserved intein residues and their corresponding block (C or N) designation. Experimental mutational studies have been carried out to further control the protein splicing reaction [3, 4]. For example, by mutating the first residue at the N-terminus (N1 block) of the intein from Cys to Ala (N1-Cys1Ala), the first step of

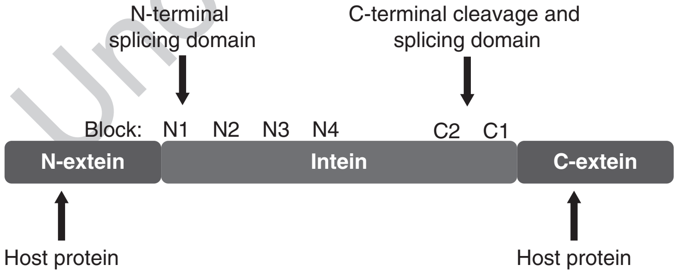

Fig. 1 Schematic intein and N- and C-exteins. Splicing motifs contain highly conserved amino acids, such as N1-Cys1, N3-His10, C2-Asp5, and C1-His7, C1-Asn8, C1-Cys+1

the splicing reaction, namely the N-terminal N–S shift,[1] is inhibited, thus isolating the C-terminal cleavage reaction [5]. Mutation schemes that control the reaction rate and/or the specific products could be exploited in many biotechnological applications such as bioseparations [6, 7], drug development [8], and molecular sensors [9, 10].

2 Methods

2.1 Computational Methodology

In order to obtain an atomic-level understanding on the reaction mechanisms as well as on the effect of mutation on the reaction barrier, we have carried out detailed quantum mechanical simulations on intein C-terminal cleavage reactions. We describe pH dependent C-terminal cleavage calculations for the *Mtu* recA intein; performed with semi-empirical, QM gas phase, QM implicit solvent, and combined QM/MM calculations [11–13]. Harnessing the C-terminal cleavage reaction may allow for an intein-based delivery device, where the reaction is triggered by a certain stimulus.

Our computational results indicate that certain mutations either inhibit or enhance specific reaction steps of the overall splicing reaction, a conclusion that is consistent with experiment. With quantum mechanical simulations, intermediate states may be isolated and studied in the context of altering the molecular triggers and inhibitors that impact protein splicing with inteins. The ability to study precursor, intermediate, and post-reaction product states is extremely useful and carried out with first principles methods.

2.2 Quantum Mechanical (QM) Methods

First principles density functional theory (DFT) [14, 15] was used to study intein C-terminal cleavage; in particular, the Becke three-parameter hybrid functional, B3LYP [16]. This hybrid method combines exchange terms from the Local Spin Density Approximation (LSDA), Hartree–Fock (HF), and Becke's (B88) exchange [17] with the correlation functionals from Lee, Yang, and Parr (LYP) [18] as well as that from the LSDA [19]. The exchange (X) and correlation (C) energy is written as E_{XC}^{B3LYP}, where

$$E_{XC}^{B3LYP} = (1-a)E_X^{LSDA} + aE_X^{HF} + b\Delta E_X^{Becke} + E_C^{LSDA} + c\Delta E_C^{LYP}, \qquad (1)$$

[1] Atoms are annotated with one letter, i.e., H = hydrogen. Amino acids are annotated with three letters, i.e., His = histidine.

and the coefficients were optimized to match extensive molecular data sets ($a = 0.20$, $b = 0.72$, and $c = 0.81$) [16]. Implemented with Gaussian code [20], this hybrid gradient-corrected method is considered one of the most accurate exchange-correlation functionals and has been used with great success in other biological systems [21, 22]. Calculations with post-Hartree–Fock Møller–Plesset perturbation theory (MP2) [23–26] were conducted to test the accuracy of the B3LYP method for this system, and the energy barrier calculations were consistent [12].

The first term in the hybrid method is E_x^{LDA}, which is the local density approximation (LDA) exchange term. E_x^{HF} is Hartree–Fock exchange integral, which is an exact quantity for electron spin exchange. Becke's B88 exchange term [17] is based on empirical results, and is written as,

$$E_x^{Becke}[\rho(\mathbf{r})] = -\beta \int d\mathbf{r}\rho(\mathbf{r})^{4/3} \frac{\alpha^2}{(1 + 6\beta \sinh^{-1} \alpha)} \tag{2}$$

where

$$\alpha = \frac{|\nabla\rho(\mathbf{r})|}{\rho(\mathbf{r})^{4/3}}.$$

Found by matching molecular data sets, β was found to be 0.0042 Hartree. Correlation functionals are from the LDA [19] and from Lee, Yang, and Parr (LYP) [18, 27], the latter based on an empirically determined model of the correlation energy of electrons in a helium atom.

Implemented with Gaussian code [20], this hybrid gradient-corrected method is considered one of the most accurate exchange-correlation functionals and has been used with great success in other biological systems [21, 22].

We have used the double-ζ basis set, 6-31G(d,p), for geometry optimizations during initial reaction path sampling [28], where the '6' represents six GTOs for core electrons and the '31' represents split GTOs for valence electrons: specifically three and one. Split-valence basis sets allow for a more accurate description of chemical bonding due to increased flexibility to fit valence electrons into molecular orbitals, and are the norm when using a Gaussian-type basis set. The '(d,p)' indicates that we are using polarization functions that allow for a shift in the wave function away from the atomic center. We have also used the triple-ζ basis set, 6-311++g(d,p), for calculations of the local minima and transition states found with the first basis set [29]. Diffuse functions for long range interactions are represented with a '+', and are especially important for anions. Basis sets of similar size are typically used for systems with similar number of electrons, and our test calculations as well as the work of others have shown these basis sets to be sufficient for similar atom types [21, 22].

2.2.1 Implicit Solvent

One method for approximating the environmental electrostatic effect is to use an implicit solvent. In this scheme, the active site is polarized by the dielectric medium which is itself polarizable. The Polarizable Continuum Model (PCM) [30] was used to simulate solvent effects in the detailed calculations. The numerical Integral Equation Formalism [31] (IEFPCM) was used because it allows for interlocking atomic spheres to represent the extent of the system in solution, which is important for protons that are in between atoms during a chemical reaction and at or around the energy barrier.

Non-dimensional dielectric constants are defined by $\varepsilon_r = \varepsilon_s/\varepsilon_0$, where ε_0 is the vacuum permittivity and ε_s is the static dielectric constant for the dielectric. For the gas phase, $\varepsilon_r = 1$. For water, $\varepsilon_r = 78.39$. Geometry optimizations were performed in implicit solvent and results are compared with gas phase calculations.

2.3 Classical Methods

Starting with the intein crystal structure for the *Mtu* RecA intein, ($\Delta\Delta$Ihh-CM, PDB code 2IN8) [32], a product protein without exteins, N- and C-terminal exteins were computationally added and then equilibrated with classical molecular dynamics (MD) simulations. The N-extein sequence consisted of Ace-Val-Val-Lys-Asn-Lys and the C-extein sequence consisted of Cys-Ser-Pro-Pro-Phe-Nme, both based on the native extein sequences [33]. Ace and Nme were capping residues for the N and C-terminal exteins, respectively. AMBER force field parameters [34] were implemented with GROMACS code [35]. MD simulations were carried out for 4 ns (0.5 ns equilibration, 3.5 ns production run) with temperature T = 298 K, pressure = 1bar, and number of water molecules = 9548 for Cys and 9549 for Met systems.

2.4 Multiscale (QM/MM) Methods

The QM/MM layering method involves treating the protein active site and critical solvent molecules with first principles methods while treating the remaining full-protein system with classical force fields [36]. The classical periodic system was trimmed down to include the protein (intein and exteins) as well as all interior waters and those exterior water molecules within a range of 7.0 Å to the protein surface (as a reference, the lone protein is roughly shaped like an oblate spheroid and approximately $25 \times 35 \times 35$ Å^3). All atoms were relaxed, and each calculation included at least 6,500 atoms. The full-protein plus solvent system, termed the real system, was treated only with the MM method. Within the real system, the active site model system was partitioned, and was treated independently by QM and MM methods. Dangling bonds that were introduced by partitioning the model system were then passivated with hydrogen atoms. With normal QM/MM energy calculations and geometry optimizations, protein and solution outside the model system was only included as a mechanical perturbation. For this reason, it is critical that the model system should include protein segments and solution molecules that are interacting electrostatically. The combined Hamiltonian may be written:

$$E^{QM:MM}_{ONIOM} = E^{QM}_{Model} - E^{MM}_{Model} + E^{MM}_{Real} \tag{3}$$

For the smaller model system, E^{QM}_{Model} is the energy calculated with quantum mechanical methods while the energy calculated by classical molecular mechanics methods is given by E^{MM}_{Model}. The real system (full protein + solvent) energy is calculated with MM methods and is given by E^{MM}_{Real}. In addition to the mechanical perturbation on the QM Hamiltonian, the electrostatic contribution from the partial charges of the MM region can be included as a perturbation on the QM Hamiltonian [37]. The QM/MM formalism has been used with success in previous work [38–40]. Typically, we report E^{QM}_{Model}, which represents the QM active site energy. The other energy terms, including the combined $E^{QM:MM}_{ONIOM}$ involves classical parameters determined for equilibrium structures that have no relevance to the energies of bond forming and breaking at transition states.

2.4.1 Charge Embedding

In addition to the mechanical perturbation on the QM Hamiltonian, the electrostatic contribution from the partial charges of the MM region can be included as a perturbation on the QM Hamiltonian. For this scheme the partial charges are those used in the MM calculation and are scaled by the default manner where atoms bonded to the inner-most four layers and atoms outside that threshold are not included [37]. Typically, we report E^{QM}_{Model}, which represents the QM active site energy. The other energy terms, including the combined $E^{QM:MM}_{ONIOM}$ involves classical parameters that have no relevance to the energies of bond forming and breaking at transition states.

2.5 Geometry Minimization

Due to the complexity of biomolecular reactions, a rigorous multidimensional search over local conformational space is essentially required although not computationally feasible for large systems [41]. Due to the time expense for each calculation, we have used the constant minimization procedure. For intermediate states along the reaction path, one coordinate is constrained while the remaining system is relaxed. The constrained internal coordinate, called the Asn cyclization distance, was the atomic distance between the Asn side chain N atom and the carbonyl C of Asn on the scissile peptide bond. In calculations with a hydronium ion (H_3O^+), the three O–H bond distances were often constrained to 0.98 Å to avoid spontaneous proton donation observed otherwise.

3 Results

3.1 Non-essential Mutation

Once splicing was inhibited, the downstream Cys residue (which was the first amino acid of the C-terminal extein or C-extein) was found to be functionally unnecessary for the C-terminal cleavage mechanism. Interestingly, Wood et al. observed that this amino acid regulated the reaction rate but did not alter the mechanism [42]. Furthermore, since the CM was found to be exceedingly reactive at low pH values, Wood et al. [42] utilized Met, which was the native N-terminus of the protein that formed the C-extein sequence, to decrease the reaction rate by an order of magnitude. In this experiment, three proteins of various sizes were contrasted with only the Cys/Met C-extein mutation: Thymidylate synthase (31.5 kDa), Hfq Protein (18 kDa), and rh aFGF (14 kDa). For these proteins, the Cys to Met mutation resulted in a decrease of the reaction rate by a factor of 12.0, 5.0, and 7.8, respectively [42,43]. Figure 2 shows a schematic of the intein precursor and products based on these results [10,44], although the exact mechanisms that govern the splicing and cleavage reactions are not understood at the atomic level. In particular, the effect of the single amino acid mutation at C + 1, flanking the conserved C1: His7-Asn8 dipeptide at the intein terminus, on the reaction rate is not understood.

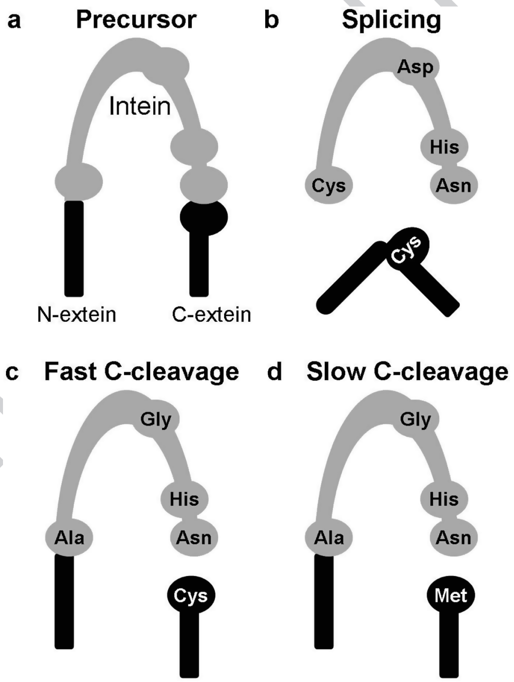

Fig. 2 Intein and extein precursor (**a**) and three possible reactions based on mutagenesis results: splicing product (**b**), and fast (**c**) and slow (**d**) C-terminal cleavage product

In order to obtain an atomic-level understanding of the effect of mutation on the reaction barrier, detailed quantum mechanical calculations on the intein C-terminal cleavage reaction have been carried out [12]. Simulations were based on both full quantum mechanical molecular analysis as well as a hybrid quantum mechanics and molecular mechanics (QM/MM) approach where the entire protein and solvent are treated classically with parameterized force fields in a molecular mechanics (MM) calculation as shown in Fig. 3a. The 53 atom C-terminal catalytic site (C1-block:

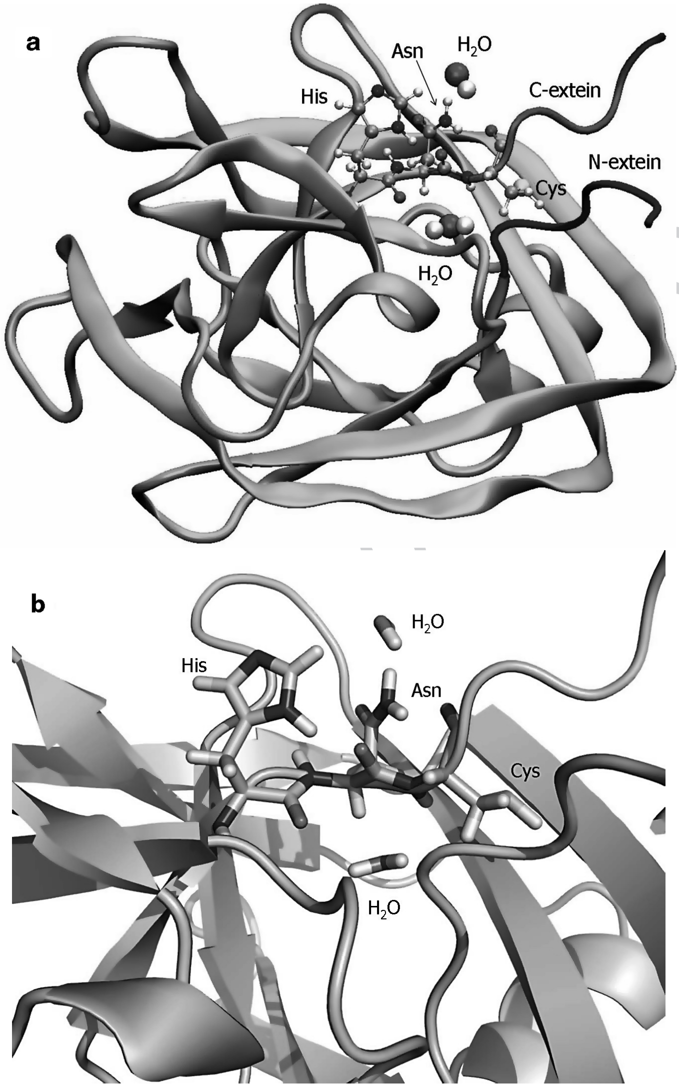

Fig. 3 The intein cleavage mutant (CM) crystal structure (PDB code 2IN8) with computationally added exteins (**a**). The C-terminal catalytic site (His–Asn–Cys + two water molecules) is highlighted (**b**)

His–Asn–Cys, or His–Asn–Xxx, where Xxx is an alternate amino acid) was treated with quantum mechanics (QM) and is shown in Fig. 3b.

The computational energy barrier was smaller for the C-terminal sequence His–Asn–Cys than for that of the His–Asn–Met mutant, consistent with experimental observations [42, 43]. The difference in energy barrier between Cys/Met residues was due to the difference in electron affinity of the amino acids. In addition to Cys and Met, several other amino acids at the first C-extein position (C + 1) were studied here. The energy barrier for C-terminal cleavage, calculated with a larger model system, is confirmed to match with that of the experiment.

3.2 Classical Protein System

Starting with the intein crystal structure for the *Mtu* recA intein, ($\Delta\Delta$Ihh-CM, PDB code 2IN8) [32], a product protein without exteins, N- and C-terminal exteins were computationally added and then equilibrated with classical molecular dynamics (MD) simulations. The N-extein sequence consisted of Ace-Val-Val-Lys-Asn-Lys and the C-extein sequence consisted of Cys-Ser-Pro-Pro-Phe-Nme, both based on the native extein sequences [33]. Ace and Nme were capping residues for the N and C-terminal exteins, respectively. AMBER force field parameters [34] were implemented with GROMACS code [35]. MD simulations were carried out for 4 ns (0.5 ns equilibration, 3.5 ns production run) with temperature T = 298K, pressure = 1bar, and number of water molecules = 9, 548 for Cys and 9,549 for Met systems.

3.3 Tripeptide Subsystem

3.3.1 Description of Model System

The tripeptide active site system (His–Asn–Cys) is highlighted in the view of the full intein crystal structure in Fig. 3b. Gas phase calculations were used to study the effect of site-directed mutagenesis (see Fig. 4). Intein crystal structures usually include a hydrogen bond between the N$^{\delta}$–H of the (penultimate) His side chain and the carbonyl O of Asn, the final amino acid of the intein [45–49]. Although the penultimate intein His residue has been previously assumed to be the proton donor for C-terminal cleavage reaction in the context of splicing [50], further inspection revealed that this was not the case for pH dependent C-terminal cleavage. For a simple proton-catalyzed reaction, there is an inverse linear rate dependence on the pH, which was observed experimentally for the C-terminal cleavage reaction [42]. Since the ability of His to act as an acid is based on its local pK$_a$ value, the expected pH-rate curve should be non-linear, specifically sigmoidal in shape, which is in contrast to the linearity observed experimentally.

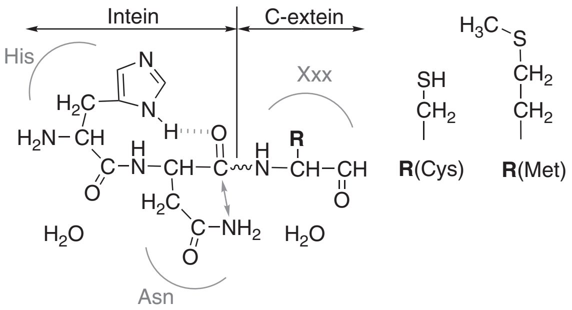

Fig. 4 The C1-block His–Asn–Xxx active site is shown. The highly conserved H-bond is shown with a dotted line, the cyclization coordinate of Asn is shown with an arrow, and the scissile peptide bond is shown with a wavy line. Side chains for Cys and Met are shown, although Ala, Val, Thr and Ser were also considered

The proposed N-protonation mechanism begins with the protonation of the peptide N by a hydronium ion (H_3O^+). This in turn causes the scissile peptide bond to elongate, and hence reduces the energy necessary for peptide bond cleavage after Asn cyclization. After Asn cyclization and aminosuccinimide formation, the extra proton passes to the cleaved C-extein N-terminus (-NH_2), which is excised and leaves with a positive charge (-NH_3^+. Although O-protonation was more energetically favorable for a generic or average peptide that was fully solvent exposed, in the case of the intein C-terminal active site, the carbonyl O was strongly hydrogen bonded to the N^δ–H of His and was also pointed inward, toward the core of the protein and away from the main body of solvent. The Asn cyclization reaction after O-protonation instead of N-protonation has been shown to require more energy and does not lead to cleavage of the peptide bond [12].

Prior to the QM/MM full protein study, the His–Asn–Cys tripeptide system (Fig. 4) was studied with an isolated gas phase reaction.[2] Certain constraints were included to ensure that the backbone structure reflects that of the protein crystal structure: both terminal backbone atoms were geometrically fixed in the crystal structure configuration, both dihedral angles are constrained to values from the crystal structure and throughout the classical molecular dynamic trajectories, and the hydrogen bond between N^δ–H of His and the carbonyl O of Asn was constrained at a distance of 1.8 Å. Without these constraints, the subsystem would likely rearrange into a structure that does not represent the intein C-terminal structure but does minimize the gas phase energy. By contrasting the effects of mutations, electronic structure properties at critical points were studied, including those at the purely quantum mechanical transition state.

[2] Gas phase energy barriers are typically higher than barriers that include electrostatic contributions such as implicit solvent calculations.

3.3.2 Energetic Results

For the N-protonation mechanism calculated with the tripeptide system, the computational energy barrier for the His–Asn–Cys system in the gas phase was 27.95 kcal/mol, in good agreement with the experimental results of ~21 kcal/mol [42]. For a system roughly 30 atoms smaller, the previous gas phase energy barrier was ~33 kcal/mol [12]. This difference indicates that even the most basic approximation of the tertiary structure is important for accurate prediction of certain reaction energy barriers, as we will see with the QM/MM reaction. Additionally, we have tested and confirmed that the hydrogen bond between N^{δ}–H of His and the carbonyl O of Asn (dashed line in Fig. 4) caused O to not accept a proton from H_3O^+. This hydrogen bond is usually found at the C-terminus of inteins and is important for reducing the possibility of proton transfer to the carbonyl O. In fact, the normally highly exothermic reaction for H_3O^+ to donate a proton to the carbonyl O atom is endothermic for cases where O is hydrogen bonded with another group [51].

Table 1 summarizes the calculated energy barriers and relative rate constants for the gas phase tripeptide system with several His–Asn–Xxx mutations. By including additional atoms, the gas phase energy barrier with Xxx = Cys (27.95 kcal/mol) was less than the previously calculated barrier for a smaller system (33 kcal/mol [12]) due to polarity and geometrical effects. The larger system used here was expected to more closely match the experiment of 21 kcal/mol, which it does, because of the additional mechanical and electronic influences of nearby protein and solvent groups.

The energy barrier of the His–Asn–Met system was 1.63 kcal/mol higher than the His–Asn–Cys system, which corresponds to a 5.83% increase in the energy barrier. When Cys was mutated to Met, the relative C-terminal reaction rate was predicted to be 0.07 as fast, or decreased by more than an order of magnitude (14.0), which is consistent with experimental results [42, 43]. Interestingly, this model predicts that Thr and Ser instead of Cys will be slightly more effective at pH-dependent C-terminal cleavage, a prediction that is consistent with the +1 position being oc-

Table 1 Tripeptide energy barriers (ΔE) for various C-extein mutations (His–Asn–Xxx), percent change (%ΔE) from His–Asn–Cys energy barrier, and expected change in reaction rate k_{rel} compared to His–Asn–Cys. Structures were geometrically optimized with the B3LYP/6-311++G(d,p) level of theory. The percent change in the energy barrier, $\%\Delta E \equiv \frac{\Delta E_{Xxx} - \Delta E_{Cys}}{\Delta E_{Xxx}} * 100\%$. Reaction rates k are relative to the His–Asn–Cys wildtype at $T = 310.15$ K (37 °C). The Arrhenius equation was used to compare the relative reaction rates between two mutants: $k = k_1/k_2 = e^{-(\Delta E_1 - \Delta E_2)/RT}$, where k_i and ΔE_i were the reaction rate and energy barrier for the i^{th} mutant, respectively; R was the gas constant and T was the temperature in Kelvin

Mutant (Xxx)	ΔE (kcal/mol)	%ΔE	k_{rel}
Cys	**27.95**	0.00	**1**
Thr	27.56	−1.39	1.88
Ser	27.75	−0.71	1.38
Ala	28.64	2.46	0.32
Val	28.97	3.64	0.19
Met	**29.58**	**5.83**	**0.07**

cupied by Cys, Thr, or Ser in nature, and will be tested in experiment. In the context of splicing, experiments have shown that Cys, Ser, and Thr are the only amino acids with the ability to complete the transesterification step of splicing [5], which is consistent because they also are the most efficient at C-terminal cleavage according to the calculations presented here.

3.3.3 Charge Analysis

Natural Populations Analysis (NPA) [52] was used to study the electron population and the partial atomic charges. Figure 5a illustrates the effect of amino acid mutation on the scissile peptide bond distance and Fig. 5b shows the sum of the NPA charges for the mutated C-extein residue, starting with the -NH at the scissile junction and including the side chain. The scissile bond distance and charge results are shown as a function of each mutant's energy barrier, and include the normal amide, the N-protonated amide, and the transition state corresponding to the pH dependent C-terminal cleavage reaction. For the neutral amide, the C–N scissile peptide bond distance was 1.3492 Å for Cys, which decreased to 1.3455 Å for Met. Although this change was extremely small, it does confirm that the amino acid side chain played a small but perceptible role in the properties of a normal peptide bond (which is well known from proton exchange experiments [53]). For the N-protonation step and then the Asn cyclization transition state, the correlation between short scissile bond distance and high energy barrier was more apparent: a shorter peptide bond implied more π-bond resonance between C and N, less π-bond resonance between C and O, and more energy was required to break the C–N bond. An elongated peptide bond implied less π bonding between C and N and less energy necessary for peptide bond cleavage [54].

A correlation between the energy barrier and the net charge can be seen (Fig. 5b), especially for the Cys/Met mutation, signifying that the residues that were able to accept more electrons exhibit a reduced energy barrier whereas the residues that were less likely or unable to accept electrons displayed an increased energy barrier.

3.4 Single Amino Acid Molecules

3.4.1 Electron Affinity and Ionization Potential Analysis

To further elucidate the effect of the mutation of the first C-extein amino acid side chain on the energy barrier, the isolated Cys and Met amino acids were studied. The electron affinities (EA) and ionization potentials (IP) for each were calculated with the B3LYP/6-311++G(d,p) level of theory. The EA for Cys, (the amount of energy gained or lost when the system goes from neutral to negatively charged), was 6.79 kcal/mol. For Met, the EA was 8.27 kcal/mol, signifying that the side chain of the gas phase Cys residue was more electronegative than for Met. The reason that Cys

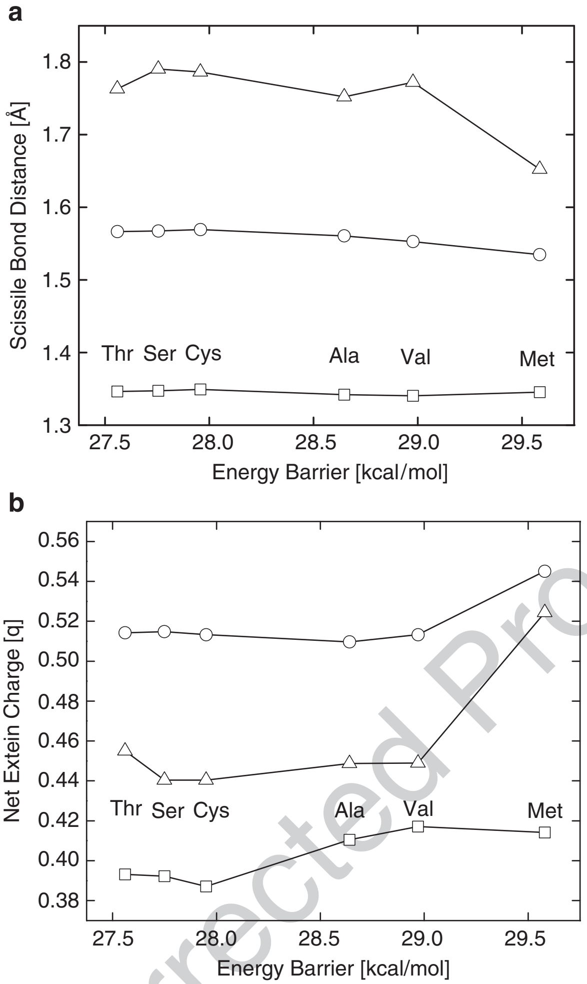

Fig. 5 Relaxed scissile peptide bond distance (**a**) and NPA charges summed for atoms on the C-extein (**b**) for the tripeptide gas phase system, His–Asn–Xxx (Xxx = Thr, Ser, Cys, Ala, Val, Met). Both the scissile bond distance and the net charge for the C-extein amino acid (Xxx) are plotted as a function of the specific mutant's energy barrier and are shown for the normal amide, (□); the N-protonated amide, (○); and the Asn cyclization transition state (△)

was more stable with charge than Met was due to the bonding for each S atom. Although each side chain contained an S atom, for Cys the S atom was bonded to one methyl group and one H atom. For Met, both bonds of the S atom were to methyl groups, hence different electron occupation properties. In changing from neutral to negatively charged, the partial charge of S for Cys changed from −0.01051 to −0.11874 units of charge, corresponding to the addition of 0.10823 electrons. For Met, the charge went from 0.16894 to 0.12532 units of charge, corresponding to the gain of only 0.04362 electrons. The S of Cys was able to accommodate more than twice the amount of delocalized electron population as compared to Met, indicating more energetic stability in the negatively charged system. The difference in ionization potential (IP) for the same isolated Cys and Met amino acids was calculated. The removal of one electron from Cys required 203.05 kcal/mol while that for Met was 191.14 kcal/mol. Combining the fact that Met was more stable when an electron was removed, and the fact that Cys was more stable when an electron was added, we conclude that the "electron pulling" and "electron pushing" properties of the first C-extein amino acid side chain must have an effect on the actual properties of the scissile peptide bond.

3.4.2 Energetic Analysis of Molecular Orbitals near the Fermi Energy

For the isolated amino acids (Thr, Ser, Cys, Ala, Val, and Met), the highest occupied molecular orbital (HOMO) for the neutrally charged system as well as the negatively charged system was compared. The difference in energy between the HOMO of the electron doped (negatively charged) and the neutral system is termed the energy gap, and is shown in Fig. 6. From this analysis of the negatively charged amino acids (geometrically optimized with neutral charge), the isolated amino acids are ranked in

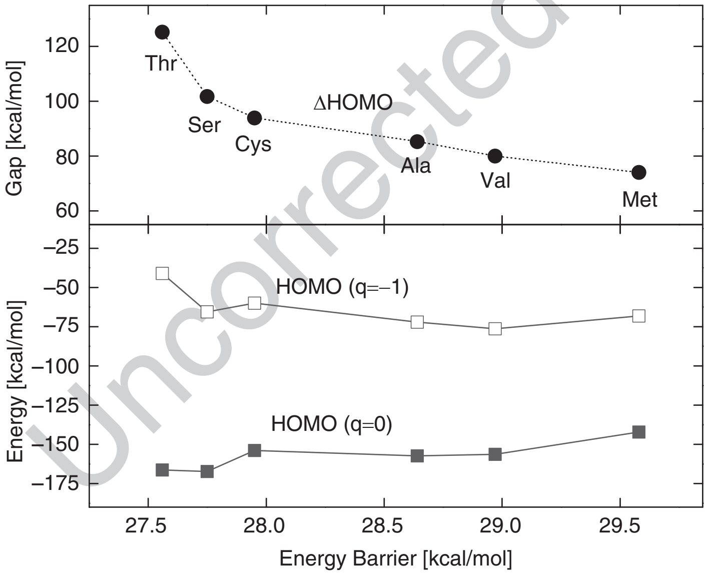

Fig. 6 Energies for the highest occupied molecular orbital (HOMO) for the neutral system (■) and the negatively charged system (□) for the isolated amino acid molecules (Thr, Ser, Cys, Ala, Val, Met), shown in order of their energy barrier found independently for the tripeptide reaction calculation. The difference between these energies is the energy gap (●) and is clearly dependent to the energy barrier for the given mutant

order of the energy barrier found when they are the mutant for the tripeptide system, and there was a clear trend in the energy gap between the neutral and negatively charged molecules. The energy gap was closely related to the electron affinity of the molecule: as the energy barrier increased for a particular mutant, the gap decreased. This single amino acid analysis is of particular interest because from the electronic structure properties of an isolated molecule representing an amino acid side chain, calculated properties such as the electron affinity, the ionization potential, and the molecular orbital energy levels may explain and perhaps predict the relative reaction rate for an unknown mutant at the first C-extein position.

The localization of the EA densities found for molecules characterized in Fig. 6 is plotted as a volumetric surface in Fig. 7, which shows the difference in electron density between the neutral (optimized geometry) and negatively charged (single point geometry) single amino acid residues (Thr, Ser, Cys, Ala, Val, and Met). The presence of electrons on the molecular side chain was observed for amino acids that are more efficient when downstream of the scissile peptide bond in intein C-terminal cleavage.

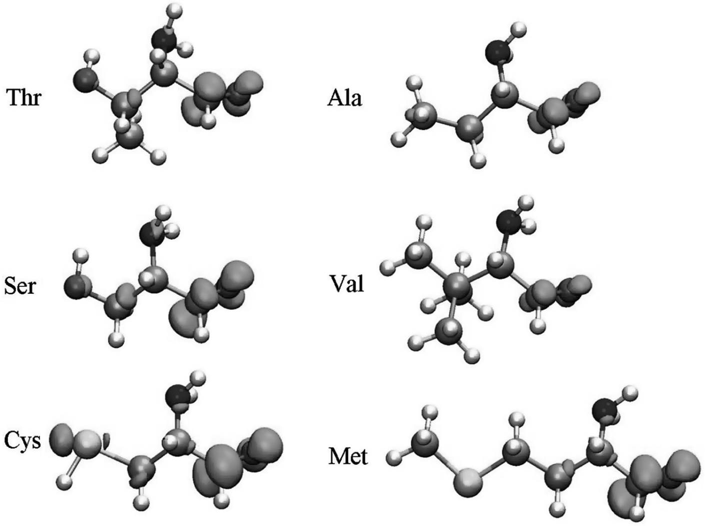

Fig. 7 The electron affinity (EA) density for single amino acid molecules (Thr, Ser, Cys, Ala, Val, and Met). The electron density surface describes the delocalization of the electron affinity when an electron is added to the system, thus going from neutral to negatively charged ($\Delta\rho$). For downstream amino acids that were efficient at C-terminal cleavage (Thr, Ser, Cys), the EA density extended to the side chain. For amino acids that were less efficient (Ala, Val, Met), the EA density remained on the peptide-like part of the molecule, and away from the side chain. Atom colors are as follows: carbon is cyan, nitrogen is blue, oxygen is white, sulfur is yellow, and hydrogen is white; the electron density surface is green [55]

3.4.3 Tripeptide Analysis

Returning to the tripeptide system shown in Fig. 4, Table 2 shows electron population analysis for orbitals with $l = 1$ angular momentum ($2s$ orbital), as well as total occupation for $l = 0, 1$ ($2s$ and $2p$ orbitals). From the analysis of target atoms belonging to the scissile peptide bond, the expected differences in electron population between Cys/Met mutants were observed. Specifically, the N atom for Met was generally more occupied with electrons than Cys, which gave it a greater negative charge.

For both mutants, the N atom showed a considerable increase of $2s$ electrons, which corresponded to C and other atoms returning σ electrons to N when the C-N bond was elongated after N-protonation. A similar situation with σ electron back-transfer to N was found for peptide bond rotation, where at the transition state of 90° the N atom lost π electrons although there was an increase in σ electrons to N [54]; this phenomenon explains why N actually became more negative as similarly seen in the present study. The $2p$ orbitals for N showed distinct differences for the Cys/Met mutations – even for the neutral ground state which was a normal amide system, a distinction that signified the side chains of adjacent amino acids were important in dictating the exact properties of the peptide bond.

For the normal amide, the charge of the peptide N for Cys was -0.616 and for Met the charge was -0.641. For the N-protonation case, the charge of N for the Cys case was -0.660, where for Met the charge was -0.710. For the transition state, the charge on N for Cys was -0.684, and for Met was -0.699. For all three cases the charge of N for Met was more negative than for Cys, which was consistent with the electron affinity calculation described previously. The side chain plays a subtle yet important role in the electrostatic environment during the cleavage reaction. By having less charge on N, the -NH_2 group is more energetically favored to leave. From this electron population analysis, differences in the electronic structure of the scis-

Table 2 Atomic orbital populations for the $2s$ and net $2p$ orbitals as well as the total electronic occupation for the peptide N atom in the gas phase tripeptide calculation. N is generally less occupied by electrons for Cys as compared to Met, which is consistent with single amino acid electron affinity results. The sum of electron occupation for the $2p_x$, $2p_y$, and $2p_z$ orbitals is written as $2p$. The NPA charge is calculated by subtracting the total electron occupation from the atomic number; a larger electron occupation signifies a more negative charge

		Occupation		
Orbital	Mutant	Neutral ground state	N-protonated	Transition state
[$2s$]	Cys	1.250	1.359	1.386
	Met	1.259	1.360	1.376
[$2p$]	Cys	4.341	4.285	4.277
	Met	4.357	4.329	4.299
Total	Cys	7.616	7.660	7.684
	Met	7.641	7.710	7.699

sile peptide bond for Cys and Met were observed, which explained why the energy barrier for Cys and Met mutants would be distinct despite an identical mechanism.

4 Reaction Analysis with QM/MM Calculations

The full protein QM/MM reaction profile was initially calculated with the QM active site region of His–Asn–Cys, and two water molecules (2346 protein atoms, 4161 water atoms, and total 53 QM atoms) [56]. Figure 8 shows the QM/MM energy barrier with and without electrostatic embedding. The energy barrier was 24.96 kcal/mol for the QM/MM calculation with geometry optimization, in excellent agreement with the 21 kcal/mol measured experimentally [42].

4.1 Effect of Mutation on Energy Barriers

The energy barrier difference for the Cys/Met mutation is of interest in the context of a QM/MM calculation, but because the Met side chain was too spatially extended to simply replace the smaller Cys side chain, additional classical MD simulations were performed (starting from the initial intein plus extein structure) but with Met at the C-extein +1 residue. Once the full protein system was equilibrated, the QM active site was partitioned to be His–Asn–Met plus the two water molecules in the same location as before (59 total QM atoms). The Asn cyclization reaction coordinate was scanned after N-protonation by H_3O^+. To compare the effect of the Met/Cys mutation directly, the smaller Cys was substituted for Met, and the geometry was again relaxed. By doing this, the change in reaction energies may be compared directly because the original protein structures were common for both Met and Cys residues.

These structures were in near total overlap, with the exception of the side chain of the (+1) amino acid, either -CH_2-SH for Cys, or -$(CH_2)_2$-S-CH_3 for Met. Using the B3LYP/6-31G(d,p) level of theory, independent reaction profiles for the Met/Cys mutation were calculated. For Met the barrier was 27.07 kcal/mol and for Cys was 26.17 kcal/mol. The His–Asn–Met QM active site (as part of the QM/MM system) had an energy barrier of 0.90 kcal/mol higher than His–Asn–Cys, which corresponded to ratio between reaction rates of $k = k_{Cys}/k_{Met} = 0.22$, in good agreement with experimental results and consistent with the tripeptide system conclusions [42, 43].

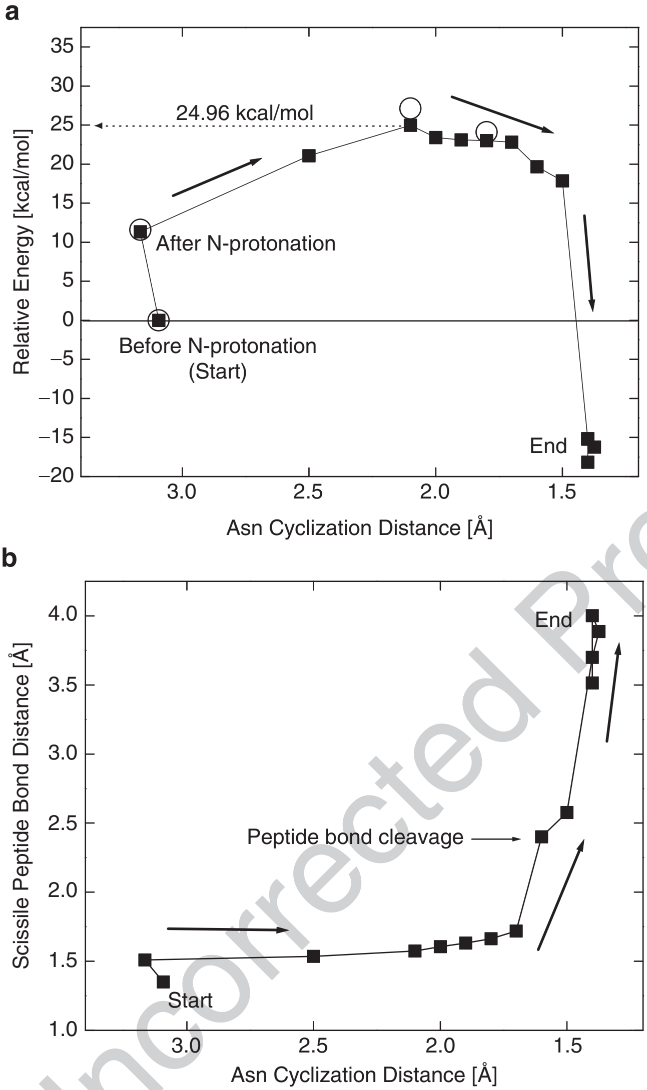

Fig. 8 Combined QM/MM reaction energy profile (**a**) and distance of the scissile peptide bond during breakage (**b**) for His–Asn–Cys plus two water QM system. QM/MM geometry optimization (■). QM/MM + charge embedding single point energies (○)

4.2 Effect of Mutation on Electron Occupation

In addition to energy barriers, the Mulliken charge [57] was calculated for critical atoms.[3] For the N atom of the scissile bond and for the ground state, the partial charge was −0.538 for Cys and for Met was −0.545. For the N-protonation state the partial charge of N was −0.609 for Cys and was −0.615 for Met. At the transition state, the charge for Cys was −0.584 and for Met was −0.598. In all cases the partial charge of the N atom for the Met mutant was more negative, which was consistent with the tripeptide results, and is explained by using the electron affinity and ionization potential for the isolated Cys and Met amino acids. When the net Mulliken charge was summed for the C-extein residue (Cys or Met) in the QM/MM context for the normal amide ground state, for Met the net charge was 0.225, and for Cys the net charge was 0.209.

Within the QM/MM system, the charge for the backbone and side chain of the first C-extein residue was added. The net charge of Cys was more negative than Met, which is in agreement with the model QM calculations described in the preceding paragraphs.

By combining model system QM calculations and full-protein QM/MM simulations, the non-mechanistic regulation of reaction rate regulation for single amino acid mutations near to the active site was confirmed, explained, and predicted. Similar methods are also useful for testing an unknown mechanism based on the correlated experimental results of kinetic data (from non-essential amino acid site-directed mutagenesis).

5 Conclusions

The C-terminal cleavage reaction and the previously proposed N-protonation mechanism were tested by increasing the QM system size by 30 atoms to at least 53 atoms. In addition, full-protein QM/MM analysis was performed. The pH dependent C-terminal cleavage reaction undergoes simple proton-catalysis by a hydronium ion that protonates the peptide N atom. The peptide bond, now resonance destabilized, is elongated and the peptide C atom is open for attack by the Asn side chain. During Asn cyclization, the peptide bond cleaves while an aminosuccinimide ring is formed. The final step involves the donation of the extra proton on the aminosuccinimide to the -NH_2 leaving group *via* water, thus making the leaving group positively charged. Our QM/MM results included the effects from the protein interior, both mechanical and electrostatic.

The "non-mechanistic" role of the first amino acid of the C-extein was confirmed. This amino acid, although not necessary for C-terminal cleavage, did have an effect on the reaction rate by about an order of magnitude, as measured by Wood et al.

[3] Natural Population Analysis (NPA) is not implemented with QM/MM at this time.

[42, 43, 58]. In this study, the precise energy barrier for C-terminal cleavage (and hence reaction rate) was shown to be dependent on the side chain of the amino acid downstream from the scissile bond. Explained by the electron occupation and partial atomic charges for each residue at the C +1 position, considerable differences that led to a distinction in energy barriers were calculated and found to be in agreement with experimentally observed reaction rates.

Acknowledgements We gratefully acknowledge Drs. Georges Belfort, Marlene Belfort and Brian Pereira for helpful scientific discussions. Funding was provided by the New York State Office of Science, Technology, and Academic Research (NYSTAR) and National Science Foundation grant CTS03-04055-NIRT. Computational resources provided by the Computational Center for Nanotechnology Innovations (CCNI).

References

1. Belfort, M., Stoddard, B., Wood, D., Derbyshire, V.: Homing Endonucleases and Inteins (Nucleic Acids and Molecular Biology). Springer, Heidelberg (2005)
2. Perler, F., Davis, E., Dean, G., Gimble, F., Jack, W., Neff, N., Noren, C., Thorner, J., Belfort, M.: Nucl. Acid Res. **22**(7), 1125 (1994)
3. Hiraga, K., Derbyshire, V., Dansereau, J., Van Roey, P., Belfort, M.: J. Mol. Biol. **354**(4), 916 (2005)
4. Shingledecker, K., Jiang, S.Q., Paulus, H.: Arch. Biochem. Biophys. **375**(1), 138 (2000)
5. Paulus, H.: Ann. Rev. Biochem. **69**(1), 447 (2000)
6. Miao, J., Wu, W., Spielmann, T., Belfort, M., Derbyshire, V., Belfort, G.: Lab Chip **5**(3), 248 (2005)
7. Banki, M., Feng, L., Wood, D.: Nat. Methods **2**(9), 659 (2005)
8. Paulus, H.: Front. Biosci. **8**, s1157 (2003)
9. Mootz, H., Muir, T.: J. Am. Chem. Soc. **124**(31), 9044 (2002)
10. Muralidharan, V., Muir, T.: Nat. Methods **3**, 429 (2006)
11. Senn, H., Thiel, W.: Angew. Chem. Int. Edit. **48**(7) (2009)
12. Shemella, P., Pereira, B., Zhang, Y., Van Roey, P., Belfort, G., Garde, S., Nayak, S.: Biophys. J. **92**(3), 847 (2007)
13. Shemella, P., Pereira, B., Van Roey, P., Belfort, G., Garde, S., Nayak, S.: Controlling C-terminal cleavage rate of an intein through extein mutation: A quantum mechanical study. (submitted) AQ3
14. Hohenberg, P., Kohn, W.: Phys. Rev. B **136**(3B), 864 (1964)
15. Kohn, W., Sham, L.: Phys. Rev. **140**(4A), 1133 (1965)
16. Becke, A.: J. Chem. Phys. **98**, 5648 (1993)
17. Becke, A.: Phys. Rev. A **38**(6), 3098 (1988)
18. Lee, C., Yang, W., Parr, R.: Phys. Rev. B **37**(2), 785 (1988)
19. Perdew, J., Wang, Y.: Phys. Rev. B **45**(23), 13244 (1992)
20. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Montgomery, Jr. J.A., Vreven, T., Kudin, K.N., Burant, J.C., Millam, J.M., Iyengar, S.S., Tomasi, J., Barone, V., Mennucci, B., Cossi, M., Scalmani, G., Rega, N., Petersson, G.A., Nakatsuji, H., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Klene, M., Li, X., Knox, J.E., Hratchian, H.P., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R., Stratmann, R.E., Yazyev, O., Austin, A.J., Cammi, R., Pomelli, C., Ochterski, J.W., Ayala, P.Y., Morokuma, K., Voth, G.A., Salvador, P., Dannenberg, J.J., Zakrzewski, V.G., Dapprich, S., Daniels, A.D., Strain, M.C., Farkas, O., Malick, D.K., Rabuck, A.D., Raghavachari, K., Foresman, J.B., Ortiz, J.V., Cui, Q., Baboul, A.G., Clifford, S., Cioslowski, J., Stefanov, B.B., Liu, G., Liashenko, A., Piskorz, P.,

Komaromi, I., Martin, R.L., Fox, D.J., Keith, T., Al-Laham, M.A., Peng, C.Y., Nanayakkara, A., Challacombe, M., Gill, P.M.W., Johnson, B., Chen, W., Wong, M.W., Gonzalez, C., Pople, J.A.: Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT (2004)

21. Zhang, X., Zhang, X., Bruice, T.: Biochemistry **44**(31), 10443 (2005)
22. Zhang, X., Bruice, T.: Proc. Natl. Acad. Sci. **103**(44), 16141 (2006)
23. Moeller, C., Plesset, M.: Phys. Rev. **46**(7), 618 (1934)
24. Head-Gordon, M., Pople, J., Frisch, M.: Chem. Phys. Lett. **153**(6), 503 (1988)
25. Frisch, M., Head-Gordon, M., Pople, J.: Chem. Phys. Lett. **166**(3), 275 (1990)
26. Frisch, M., Head-Gordon, M., Pople, J.: Chem. Phys. Lett. **166**(3), 281 (1990)
27. Miehlich, B., Savin, A., Stoll, H., Preuss, H.: Chem. Phys. Lett. **157**(3), 200 (1989)
28. Hehre, W., Ditchfield, R., Pople, J.: J. Chem. Phys. **56**(5), 2257 (1972)
29. McLean, A., Chandler, G.: J. Chem. Phys. **72**(10), 5639 (2006)
30. Miertus, S., Scrocco, E., Tomasi, J.: Chem. Phys. **55**(11), 117 (1981)
31. Cances, M., Mennucci, B., Tomasi, J.: J. Chem. Phys. **107**(8), 3032 (1997)
32. Van Roey, P., Pereira, B., Li, Z., Hiraga, K., Belfort, M., Derbyshire, V.: J. Mol. Biol. **367**(1), 162 (2007)
33. Davis, E., Sedgwick, S., Colston, M.: J. Bacteriol. **173**(18), 5653 (1991)
34. Cornell, W., Cieplak, P., Bayly, C., Gould, I., Merz, K., Ferguson, D., Spellmeyer, D., Fox, T., Caldwell, J., Kollman, P.: J. Am. Chem. Soc. **117**(19), 5179 (1995)
35. Van der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A., Berendsen, H.: J. Comput. Chem. **26**(16), 1701 (2005)
36. Maseras, F., Morokuma, K.: J. Comput. Chem. **16**(9), 1170 (1995)
37. Vreven, T., Morokuma, K., Farkas, O., Schlegel, H., Frisch, M.: J. Comput. Chem. **24**(6), 760 (2003)
38. Gao, J., Truhlar, D.: Ann. Rev. Phys. Chem. **53**, 467 (2002)
39. Cui, Q., Elstner, M., Karplus, M.: J. Phys. Chem. B **106**(10), 2721 (2002)
40. Torrent, M., Vreven, T., Musaev, D., Morokuma, K., Farkas, O., Schlegel, H.: J. Am. Chem. Soc. **124**(2), 192 (2002)
41. Dellago, C., Bolhuis, P., Csajka, F., Chandler, D.: J. Chem. Phys. **108**(5), 1964 (1998)
42. Wood, D., Derbyshire, V., Wu, W., Chartrain, M., Belfort, M., Belfort, G.: Biotechnol. Progr. **16**(6), 1055 (2000)
43. Wood, D.: Generation and application of a self-cleaving protein linker for use in single-step affinity fusion based protein purification. Ph.D. thesis, Rensselaer Polytechnic Institute (2000)
44. Wu, W., Wood, D., Belfort, G., Derbyshire, V., Belfort, M.: Nucl. Acid Res. **30**(22), 4864 (2002)
45. Mizutani, R., Nogami, S., Kawasaki, M., Ohya, Y., Anraku, Y., Satow, Y.: J. Mol. Biol. **316**(4), 919 (2002)
46. Poland, B., Xu, M., Quiocho, F.: J. Biol. Chem. **275**(22), 16408 (2000)
47. Klabunde, T., Sharma, S., Telenti, A., Jacobs, W., Sacchettini, J.: Nat. Struct. Biol. **5**(1), 31 (1998)
48. Duan, X., Gimble, F., Quiocho, F.: Cell **89**(4), 555 (1997)
49. Ichiyanagi, K., Ishino, Y., Ariyoshi, M., Komori, K., Morikawa, K.: J. Mol. Biol. **300**(4), 889 (2000)
50. Ding, Y., Xu, M., Ghosh, I., Chen, X., Ferrandon, S., Lesage, G., Rao, Z.: J. Biol. Chem. **278**(40), 39133 (2003)
51. AQ4 Shemella, P., Nayak, S.: (unpublished)
52. Reed, A., Curtiss, L., Weinhold, F.: Chem. Rev. **88**(6), 899 (1988)
53. Bai, Y., Milne, J., Mayne, L., Englander, S.: Prot. Struct. Funct. Genet. **17**(1), 75 (1993)
54. Milner-White, E.: Prot. Sci. **6**(11), 2477 (1997)
14. Humphrey, W., Dalke, A., Schulten, K.: J. Mol. Graph. **14**(1), 33 (1996)
56. Pereira, B., Jain, S., Garde, S.: J. Chem. Phys. **124**, 074704 (2006)
57. Mulliken, R.: J. Chem. Phys. **23**(10), 1833 (1955)
58. Wood, D., Wu, W., Belfort, G., Derbyshire, V., Belfort, M.: Nat. Biotechnol. **17**(9), 889 (1999)

AUTHOR QUERIES

AQ1. Please provide keywords.
AQ2. Please sepcify the corresponding author.
AQ3. Please update Shemella et al.
AQ4. Please provide year for Ref. Shemella and Nayak.